

日 本 国 特 許 庁  
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出 願 年 月 日

Date of Application:

2002年 9月13日

出 願 番 号

Application Number:

特願2002-269193

[ ST.10/C ]:

[ JP2002-269193 ]

出 願 人

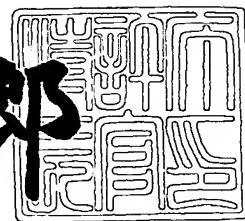
Applicant(s):

富士ゼロックス株式会社

2003年 5月16日

特 許 庁 長 官  
Commissioner,  
Japan Patent Office

太田 信一郎



出証番号 出証特2003-3036245

【書類名】 特許願

【整理番号】 FE02-01294

【あて先】 特許庁長官 殿

【提出日】 平成14年 9月13日

【国際特許分類】 G06F 17/00

【発明者】

    【住所又は居所】 神奈川県足柄上郡中井町境4 3 0 グリーンテクなかい  
                        富士ゼロックス株式会社内

    【氏名】 劉 紹明

【特許出願人】

    【識別番号】 000005496

    【氏名又は名称】 富士ゼロックス株式会社

【代理人】

    【識別番号】 100098132

    【弁理士】

    【氏名又は名称】 守山 辰雄

【手数料の表示】

    【予納台帳番号】 035873

    【納付金額】 21,000円

【提出物件の目録】

    【物件名】 明細書 1

    【物件名】 図面 1

    【物件名】 要約書 1

    【包括委任状番号】 9606109

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 テキスト文比較装置

【特許請求の範囲】

【請求項 1】 テキスト文間の比較を行うテキスト文比較装置において、  
比較対象となるテキスト文の全体の構造と意味をグラフ理論上の根がある木で  
表現する木表現手段と、  
木表現手段により表現される木の各頂点に単語情報を付与するとともに各辺に  
単語間の係り受け関係情報である格情報を付与する情報付与手段と、  
頂点の間の対応関係及び辺の間の対応関係に基づく木の間の距離を定義する木  
間距離定義手段と、  
木間距離定義手段により定義される木の間の距離を求める木間距離取得手段と  
、  
木の間の距離をテキスト文間の相違を表す距離に適用する木間距離適用手段と  
、  
木間距離適用手段による適用に基づいて、比較対象となるテキスト文間の距離  
を求めるテキスト文間距離取得手段と、  
を備えたことを特徴とするテキスト文比較装置。

【請求項 2】 請求項 1 に記載のテキスト文比較装置において、  
木間距離定義手段は、木の各頂点及び各辺にラベルを付けるラベル付け手段と  
、  
木の各頂点及び各辺に番号を付ける番号付け手段と、  
頂点間の対応関係及び辺間の対応関係に基づく木の間の写像条件を設定する木  
間写像条件設定手段と、  
木間写像条件設定手段により設定される木の間の写像条件に基づいて木の間の  
写像を行う木間写像手段と、  
頂点間の対応関係及び辺間の対応関係に基づく森の間の写像を行う森間写像手  
段と、  
写像の重みを設定する写像重み設定手段と、  
森間の写像と写像の重みに基づく森間の距離を定義する森間距離定義手段と、

木間の写像と写像の重みに基づく木間の距離を定義する距離定義手段と、を有する、

ことを特徴とするテキスト文比較装置。

【請求項3】 請求項1又は請求項2に記載のテキスト文比較装置において

木表現手段は、テキスト文全体の構造と意味をグラフ理論上の根があり順序がある木で表現する、

ことを特徴とするテキスト文比較装置。

【請求項4】 請求項1又は請求項2に記載のテキスト文比較装置において

木表現手段は、テキスト文全体の構造と意味をグラフ理論上の根があり順序がない木で表現する、

ことを特徴とするテキスト文比較装置。

【請求項5】 請求項2に記載のテキスト文比較装置において、

木表現手段は、テキスト文全体の構造と意味をグラフ理論上の根があり順序がある木で表現し、

森間写像手段は、頂点間の対応関係及び辺間の対応関係に基づく順序がある森の間の写像を行い、

森間距離定義手段は、順序がある森の間の写像と写像の重みに基づく順序がある森の間の距離を定義する、

ことを特徴とするテキスト文比較装置。

【請求項6】 請求項2に記載のテキスト文比較装置において、

木表現手段は、テキスト文全体の構造と意味をグラフ理論上の根があり順序がない木で表現する

森間写像手段は、頂点間の対応関係及び辺間の対応関係に基づく順序がない森の間の写像を行い、

森間距離定義手段は、順序がない森の間の写像と写像の重みに基づく順序がない森の間の距離を定義する、

ことを特徴とするテキスト文比較装置。

【請求項 7】 請求項 1 乃至請求項 6 のいずれか 1 項に記載のテキスト文比較装置において、

木間距離適用手段は、単語の写像と木の頂点の写像とを対応させる単語写像対応手段と、

格の写像と木の辺の写像とを対応させる格写像対応手段と、

単語の写像の重みを設定する単語写像重み設定手段と、

格の写像の重みを設定する格写像重み設定手段と、

単語写像重みと木の頂点の写像重みとを対応させる単語写像重み対応手段と、

格写像重みと木の辺の写像重みとを対応させる格写像重み対応手段と、を有す

る、

ことを特徴とするテキスト文比較装置。

【請求項 8】 請求項 7 に記載のテキスト文比較装置において、

単語写像重み設定手段は、置換と脱落と挿入について、単語の写像の重みを設定し、

格写像重み設定手段は、置換と脱落と挿入について、格の写像の重みを設定し

単語写像重み対応手段は、単語置換重みと木の頂点間の置換重みとを対応させる単語置換重み対応手段と、単語脱落重みと木の頂点の脱落重みとを対応させる単語脱落重み対応手段と、単語挿入重みと木の頂点の挿入重みとを対応させる単語挿入重み対応手段と、を有し、

格写像重み対応手段は、格置換重みと木の辺の間の置換重みとを対応させる格置換重み対応手段と、格脱落重みと木の辺の脱落重みとを対応させる格脱落重み対応手段と、格挿入重みと木の辺の挿入重みとを対応させる格挿入重み対応手段と、を有する、

ことを特徴とするテキスト文比較装置。

【請求項 9】 請求項 7 又は請求項 8 に記載のテキスト文比較装置において

単語写像重み設定手段は、木の間の写像において 2 つの頂点が写像した場合に各頂点に格納されている単語の間の置換重みを設定する単語置換重み設定手段と

木の間の写像において頂点が写像できなく脱落された場合に頂点に格納されている単語の脱落重みを設定する単語脱落重み設定手段と、

木の間の写像において頂点が写像できなく挿入された場合に頂点に格納されている単語の挿入重みを設定する単語挿入重み設定手段と、

単語置換重みと単語脱落重みと単語挿入重みとの間の関係を設定する単語写像重み関係設定手段と、を有する、

ことを特徴とするテキスト文比較装置。

【請求項10】 請求項7乃至請求項9のいずれか1項に記載のテキスト文比較装置において、

格写像重み設定手段は、木の間の写像において2つの辺が写像した場合に各辺に格納されている格の間の置換重みを設定する格置換重み設定手段と、

木の間の写像において辺が写像できなく脱落された場合に辺に格納されている格の脱落重みを設定する格脱落重み設定手段と、

木の間の写像において辺が写像できなく挿入された場合に辺に格納されている格の挿入重みを設定する格挿入重み設定手段と、

格置換重みと格脱落重みと格挿入重みとの間の関係を設定する格写像重み関係設定手段と、を有する、

ことを特徴とするテキスト文比較装置。

【請求項11】 請求項9に記載のテキスト文比較装置において、

単語置換重み設定手段は、2つの単語が同一の単語である場合には単語置換重みをゼロと設定し、2つの単語が異なる場合には単語置換重みを正の定数と設定する、

ことを特徴とするテキスト文比較装置。

【請求項12】 請求項9に記載のテキスト文比較装置において、

単語置換重み設定手段は、単語置換重みを2つの単語の間の距離の値と設定する、

ことを特徴とするテキスト文比較装置。

【請求項13】 請求項9、請求項11、又は請求項12のいずれか1項に

記載のテキスト文比較装置において、

単語脱落重み設定手段は、単語の脱落重みを定数と設定する、  
ことを特徴とするテキスト文比較装置。

【請求項 14】 請求項 9、請求項 11、又は請求項 12 のいずれか 1 項に  
記載のテキスト文比較装置において、

単語脱落重み設定手段は、単語の脱落重みを単語の品詞に基づいて設定する、  
ことを特徴とするテキスト文比較装置。

【請求項 15】 請求項 9 又は請求項 11 乃至請求項 14 のいずれか 1 項に  
記載のテキスト文比較装置において、

単語挿入重み設定手段は、単語の挿入重みを定数と設定する、  
ことを特徴とするテキスト文比較装置。

【請求項 16】 請求項 9 又は請求項 11 乃至請求項 14 のいずれか 1 項に  
記載のテキスト文比較装置において、

単語挿入重み設定手段は、単語の挿入重みを単語の品詞に基づいて設定する、  
ことを特徴とするテキスト文比較装置。

【請求項 17】 請求項 9 又は請求項 11 乃至請求項 16 のいずれか 1 項に  
記載のテキスト文比較装置において、

単語写像重み関係設定手段は、「単語脱落重み+単語挿入重み>単語置換重み  
」という関係を設定する、

ことを特徴とするテキスト文比較装置。

【請求項 18】 請求項 10 に記載のテキスト文比較装置において、  
格置換重み設定手段は、2つの格が同一の格である場合には格置換重みをゼロ  
と設定し、2つの格が異なる場合には格置換重みを正の定数と設定する、

ことを特徴とするテキスト文比較装置。

【請求項 19】 請求項 10 に記載のテキスト文比較装置において、  
格置換重み設定手段は、全ての格を複数のカテゴリに分類する格分類手段と、  
格のカテゴリ間の置換重みを設定する格カテゴリ間置換重み設定手段と、を有し  
、格置換重みを2つの格が属しているカテゴリ間の置換重みと設定する、

ことを特徴とするテキスト文比較装置。

【請求項 20】 請求項 10、請求項 18、又は請求項 19 のいずれか 1 項に記載のテキスト文比較装置において、

格脱落重み設定手段は、格脱落重みを定数と設定する、  
ことを特徴とするテキスト文比較装置。

【請求項 21】 請求項 10、請求項 18、又は請求項 19 のいずれか 1 項に記載のテキスト文比較装置において、

格脱落重み設定手段は、格脱落重みを格の種類に基づいて設定する、  
ことを特徴とするテキスト文比較装置。

【請求項 22】 請求項 10 又は請求項 18 乃至請求項 21 のいずれか 1 項に記載のテキスト文比較装置において、

格挿入重み設定手段は、格挿入重みを定数と設定する、  
ことを特徴とするテキスト文比較装置。

【請求項 23】 請求項 10 又は請求項 18 乃至請求項 21 のいずれか 1 項に記載のテキスト文比較装置において、

格挿入重み設定手段は、格挿入重みを格の種類に基づいて設定する、  
ことを特徴とするテキスト文比較装置。

【請求項 24】 請求項 10 又は請求項 18 乃至請求項 23 のいずれか 1 項に記載のテキスト文比較装置において、

格写像重み関係設定手段は、「格脱落重み+格挿入重み>格置換重み」という関係を設定する、

ことを特徴とするテキスト文比較装置。

【請求項 25】 請求項 1 乃至請求項 24 のいずれか 1 項に記載のテキスト文比較装置において、

テキスト文間距離取得手段は、木間距離取得手段により求められる距離の値をテキスト文間の距離とする、

ことを特徴とするテキスト文比較装置。

【請求項 26】 請求項 1 乃至請求項 24 のいずれか 1 項に記載のテキスト文比較装置において、

テキスト文間距離取得手段は、木間距離取得手段により求められる距離の値を



比較対象となる木の頂点数の和で割り算した結果をテキスト文間の距離とする、  
ことを特徴とするテキスト文比較装置。

【請求項 27】 テキスト文間の比較を行うテキスト文比較方法において、  
比較対象となるテキスト文の全体の構造と意味をグラフ理論上の根がある木で  
表現し、

表現される木の各頂点に単語情報を付与するとともに各辺に単語間の係り受け  
関係情報である格情報を付与し、

頂点の間の対応関係及び辺の間の対応関係に基づく木の間の距離の定義に基づ  
いて、比較対象となるテキスト文の木間の距離を求め、

木の間の距離をテキスト文間の相違を表す距離に適用して、比較対象となるテ  
キスト文間の距離を求める、

ことを特徴とするテキスト文比較方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、例えばコンピュータを利用して、テキスト文間の意味内容の相違を  
比較する装置及び方法に関し、特に、高精度に実時間で比較を行う装置及び方法  
に関する。

【0002】

【従来の技術】

IT 技術、特に高速インターネット・モバイル技術の飛躍的な発展により、大  
量の情報が誰でも、どこでも、いつでも利用可能になったが、逆に、情報洪水と  
言われる現象が起き、ユーザが真に必要な情報を取り出すことが困難になってき  
ている。利用者がどのような状況にあっても常に適切な情報を得ることができる  
世界を実現するために、情報洪水の中からユーザにとって真に価値ある情報を抽  
出・再構成することが必要である。

【0003】

ここで、ドキュメントの意味内容の比較や意味内容によるテキスト文書の分類  
やユーザの情報探索意図の理解に関する技術は重要である。また、ドキュメント

の意味内容の比較や意味内容によるテキスト文書の分類やユーザの情報探索意図の理解を実現するためには、自然言語処理などを利用した意味の類似性判定が欠かせないものである。

【0004】

この分野では、テキスト文間の類似性を測る技術が幾つか提案されているが、文のローカルな情報、例えば、文に出現している単語情報や、単語間の係り受け情報を利用したものが主流であって、テキスト文の意味内容の評価尺度としては適用しにくく、ドキュメントの意味内容の比較及びユーザの情報探索意図の理解という目標の実現にはつながらなかった。

【0005】

最近、テキスト文を意味解析してグラフで表現し、グラフ表現の基に経験的類似度を測る手法が提案されたが、提案された類似度には、構造的な変化を考慮していないものもあるし、類似度の定義とテキスト文の意味内容の相違との関係が明確ではないものもあった。

【0006】

また、本発明と関連する従来技術の例として、以下のようなものがあった。

【0007】

【非特許文献1】

原田、水野、論文“EDRを用いた日本語意味解析システムSAGE”、「人口知能学会論文誌」、2001年、16(1)、p. 85-93

【非特許文献2】

相澤彰子、論文“語と文書の共起に基づく特徴量の定義と適用”、「自然言語処理」、2000年3月、136-4

【非特許文献3】

馬青、論文“日本語名詞の意味マップの自己組織化”、「情報処理学会論文誌」、2001年、第42巻、第10号

【0008】

【発明が解決しようとする課題】

上記で述べたように、従来では、テキスト文間の意味内容の類似度を比較する

性能は未だに不十分なものであり、従来提案された類似度はテキスト文間の意味内容の相違の説明につながりにくいといった問題があった。

## 【0009】

本発明は、上記のような従来の事情に鑑みなされたもので、テキスト文間の意味内容の相違を高精度に実時間で比較することができる装置や方法を提供することを目的とする。更に具体的には、本発明では、例えば、ドキュメントの意味内容の比較や、意味内容によるテキスト文書の分類や、ユーザの情報探索意図の理解を実現するために、テキスト文間の意味内容の相違を測ることができる距離尺度を数学的に定義して、当該距離尺度を実時間内で求めることを実現する。

## 【0010】

## 【課題を解決するための手段】

上記目的を達成するため、本発明に係るテキスト文比較装置では、次のようにして、テキスト文間の比較を行う。

すなわち、木表現手段が、比較対象となるテキスト文をグラフ理論上の根がある木で表現する。情報付与手段が、木表現手段により表現される木の各頂点に単語情報を付与するとともに各辺に単語間の係り受け関係情報である格情報を付与する。木間距離定義手段が、頂点の間の対応関係及び辺の間の対応関係に基づく木の間の距離を定義する。木間距離取得手段が、木間距離定義手段により定義される木の間の距離を求める。木間距離適用手段が、木間の距離をテキスト文間の相違（或いは、類似）を表す距離に適用する。テキスト文間距離取得手段が、木間距離適用手段による適用に基づいて、比較対象となるテキスト文間の距離を求める。

## 【0011】

従って、比較対象となる2つのテキスト文について、テキスト文全体の構造と意味をグラフ理論上の根がある木で表現し、2つの木の間の距離を適用して求められるこれら2つのテキスト文間の距離に基づいて、これら2つのテキスト文間の意味的な相違を検討することができるため、テキスト文間の比較を高精度に実時間で行うことができる。

## 【0012】

ここで、本発明では、グラフ理論上の木の間距離をテキスト文の比較に適用しており、テキスト文に含まれる単語の情報や格の情報ばかりでなく、テキスト文の構造についても考慮している。また、本発明では、木の頂点に単語情報を付与するとともに、木の辺に格情報を付与している。

また、根があり順序がある木を用いるか或いは根があり順序がない木を用いるかによって、テキスト文間の距離を2種類に分けることができ、応用上で計算速度及び比較精度によって任意に選択することができる。

#### 【0013】

なお、本明細書では、グラフ理論上の根があり順序がある木をR O 木 (R o o t e d a n d O r d e r e d t r e e) と言い、グラフ理論上の根があり順序がない木をR 木 (R o o t e d t r e e) と言う。

また、R O 木とR 木とを比較すると、R O 木の方がR 木と比べて計算が簡易である一方、R 木の方がR O 木と比べて一般的に精度がよい。

#### 【0014】

また、単語情報としては、種々な情報が用いられてもよく、例えば、単語属性の情報が含まれてもよい。単語属性の情報としては、例えば、形態素解析により得られる品詞の情報などを用いることができ、また、動詞の場合には変形の情報などを用いることができる。

また、単語間の係り受けの種類が格に相当する。

また、単語の情報及び格の情報は、例えば、テキスト文を意味解析して求められる。

#### 【0015】

また、R 木間の写像条件としては、例えば、頂点間の写像と辺間の写像に対して、「1対1写像であり、親子関係（上下関係）を保存し、構造を保存し、頂点間の写像と辺間の写像が交差しない」などの条件が用いられる。

また、R O 木間の写像条件としては、例えば、頂点間の写像と辺間の写像に対して、「1対1写像であり、親子関係（上下関係）を保存し、兄弟間に左右の関係を保存し、構造を保存し、頂点間の写像と辺間の写像が交差しない」などの条件が用いられる。

## 【 0 0 1 6 】

また、木Aから木Bへ写像するとき、例えば、木Aの頂点が木Bの頂点に写像する場合は頂点の置換に相当し、木Aにあって写像できない頂点が脱落に相当し、木Bにあって写像できない頂点が挿入に相当する。

また、木と木の間の距離としては、例えば、1本の木を他の1本の木へ変換するときの重みの和（写像重みの和）の最小値が用いられる。また、このような木と木の間の距離には、暗黙的に、森と森の間の距離が含まれる。

## 【 0 0 1 7 】

また、RO木或いはR木の各頂点に番号を付ける仕方としては、例えば、深さ優先探索により各頂点に番号を増大順で付けて、番号が大きい頂点から順に計算を行う仕方が用いられる。具体的には、動的計画法を用いて、一番下側の部分木から上側の部分木へと順に計算を行っていく。

また、ラベルは、情報を記憶するためのものである。

## 【 0 0 1 8 】

以下で、更に、本発明の構成例を示す。

(1) テキスト文間の意味内容を測る距離を求めるテキスト文間の意味内容比較装置において、テキスト文全体の構造と意味をグラフ理論上のRO木或いはR木で表現する手段と、RO木或いはR木の各頂点と各辺にそれぞれ単語情報と関連する単語間の係り受け関係（格）情報を付与する手段と、頂点間及び辺間の対応関係に基づくRO木間或いはR木間の距離を定義する手段と、定義したRO木間或いはR木間の距離を求める手段と、RO木間或いはR木間の距離をテキスト文間の意味的な相違を比較する距離に適用する手段と、テキスト文間の距離を求める手段を備える。

## 【 0 0 1 9 】

(2) 頂点間及び辺間の対応関係に基づくRO木間或いはR木間の距離を定義する手段は、RO木或いはR木の各頂点と各辺にそれぞれラベルを付けるラベル付け手段と、RO木或いはR木の各頂点と各辺にそれぞれ番号を付ける番号付け手段と、頂点間と辺間の対応関係に基づくRO木間の写像条件と、頂点間と辺間の対応関係に基づくR木間の写像条件と、頂点間と辺間の対応関係に基づくRO

木間の写像手段と、頂点間と辺間の対応関係に基づくR木間の写像手段と、頂点間と辺間の対応関係に基づく順序がある森の間の写像を行う写像手段と、頂点間と辺間の対応関係に基づく順序がない森の間の写像を行う写像手段と、これらの写像手段による写像の重みを定義する写像重み設定手段と、順序がある森の間の写像手段と写像重み設定手段に基づく順序がある森の間の距離を定義する手段と、順序がない森の間の写像手段と写像重み設定手段に基づく順序がない森の間の距離を定義する手段と、RO木間の写像手段と写像重み設定手段に基づくRO木間の距離を定義する手段と、R木間の写像手段と写像重み設定手段に基づくR木間の距離を定義する手段を有する。

## 【 0 0 2 0 】

(3) RO木間或いはR木間の距離をテキスト文間の意味的な相違を測る距離に適用する手段は、単語間の写像をRO木或いはR木の頂点間の写像に対応させる手段と、格間の写像をRO木或いはR木の辺間の写像に対応させる手段と、単語間の写像重みをRO木或いはR木の頂点間の写像重みに対応させる手段と、格間の写像重みをRO木或いはR木の辺間の写像重みに対応させる手段と、単語の写像重みを設定する手段と、格の写像重みを設定する手段を有する。

## 【 0 0 2 1 】

(4) テキスト文間の距離を求める手段は、RO木間或いはR木間の距離を求める手段で求められた距離値をテキスト文間の距離とする。

(5) テキスト文間の距離を求める手段は、RO木間或いはR木間の距離を求める手段で求められた距離値と2つのRO木或いはR木の頂点数の和とを割り算した結果をテキスト文間の距離尺度とする。

## 【 0 0 2 2 】

(6) 単語間の写像重みを設定する手段は、RO木間或いはR木間の写像において、2つの頂点が写像した場合には、各頂点に格納されている単語間の置換重みを設定する手段と、頂点が写像できなく、脱落された場合には、頂点に格納されている単語の脱落重みを設定する手段と、頂点が写像できなく、挿入された場合には、頂点に格納されている単語の挿入重みを設定する手段と、単語置換重みと単語脱落重みと単語挿入重みとの間の関係を設ける手段を有する。

【0023】

(7) 格の写像重みを設定する手段は、RO木間或いはR木間の写像において、2つの辺が写像した場合には、各辺に格納されている格間の格置換重みを設定する手段と、辺が写像できなく、脱落された場合には、辺に格納されている格の脱落重みを設定する手段と、辺が写像できなく、挿入された場合には、辺に格納されている格の挿入重みを設定する手段と、格置換重みと格脱落重みと格挿入重みとの間の関係を設ける手段を有する。

【0024】

(8) 単語置換重みを設定する手段は、2つの単語が同一の単語である場合には単語の置換重みをゼロと設定する手段、2つの単語が異なる場合には単語の置換重みを正の定数と設定する手段を有する。

(9) 単語置換重みを設定する手段は、単語の置換重みを単語間の距離と設定する。

【0025】

(10) 単語脱落重みを設定する手段は、単語の脱落重みを定数と設定する。

(11) 単語脱落重みを設定する手段は、単語の品詞によって単語の脱落重みを求める。

(12) 単語挿入重みを設定する手段は、単語の挿入重みを定数と設定する。

(13) 単語挿入重みを設定する手段は、単語の品詞によって単語の挿入重みを求める。

(14) 単語の置換重み、脱落重み、挿入重みの間の関係を設ける手段は、「単語脱落重み+単語挿入重み>単語置換重み」を満たす関係を設ける。

【0026】

(15) 格置換重みを設定する手段は、2つの格が同一の格である場合には格の置換重みをゼロと設定する手段と、2つの格が異なる場合には格の置換重みを正の定数と設定する手段を有する。

(16) 格置換重みを設定する手段は、全ての格を複数であるN個のカテゴリに分類する手段と、格カテゴリ間の置換重みを設定する手段と、格間の置換重みを、2つの格が属しているカテゴリ間の置換重みとする手段を有する。

【0027】

(17) 格脱落重みを設定する手段は、格の脱落重みを定数と設定する。

(18) 格脱落重みを設定する手段は、格の種類によって格の脱落重みを求める。

(19) 格挿入重みを設定する手段は、格の挿入重みを定数と設定する。

(20) 格挿入重みを設定する手段は、格の種類によって格の挿入重みを求める。

(21) 格の置換重み、脱落重み、挿入重みの間の関係を設ける手段は、「格脱落重み+格挿入重み>格置換重み」を満たす関係を設ける。

【0028】

(22) テキスト文間の意味内容を測る距離を求めるテキスト文間の意味内容比較方法において、テキスト文全体の構造と意味をグラフ理論上のRO木或いはR木で表現し、RO木或いはR木の各頂点と各辺にそれぞれ単語情報と単語間の係り受け関係(格)情報を格納し、頂点間及び辺間の対応関係に基づくRO木間或いはR木間の距離を定義した内容に基づいて、定義したRO木間或いはR木間の距離を求め、RO木間或いはR木間の距離をテキスト文間の意味的な相違を比較する距離に適用して、テキスト文間の距離を求める。

【0029】

【発明の実施の形態】

本発明に係る一実施例を図面を参照して説明する。

図1には、本発明の一実施例に係るテキスト文間の意味内容の比較装置(テキスト文比較装置)の実施の一形態を示してあり、当該装置は本発明の一実施例に係るテキスト文の意味内容の比較方法を実施する。

【0030】

同図に示したテキスト文比較装置には、外部記憶装置1と、テキスト文の形態素を抽出する形態素解析部2と、テキスト文の意味を解析する意味解析部3と、意味解析部3で解析した結果をグラフ理論上のRO木或いはR木に変換する木構造変換部4と、2つの単語を置換するときの単語置換重みと単語を脱落するときの単語脱落重みと単語を挿入するときの単語挿入重みを求める単語写像重み計算



部5と、2つの格を置換するときの格置換重みと格を脱落するときの格脱落重みと格を挿入するときの格挿入重みを求める格写像重み計算部6と、R〇木間或いはR木間の距離を計算する距離計算部7と、テキスト文間の意味内容の相違を求める意味内容比較部8と、例えばメモリから構成される記憶部9と、複数のメモリ10～18が備えられている。

## 【0031】

なお、外部記録装置1には、テキスト文のデータが格納されている。

また、メモリ10とメモリ11は、外部記憶装置1から読み出した2つのテキスト文のデータをそれぞれ記憶する。メモリ12とメモリ13は、2つのテキスト文のそれぞれについて、形態素解析部2による解析結果を記憶する。メモリ14とメモリ15は、2つのテキスト文のそれぞれについて、意味解析部3による意味解析の結果を記憶する。メモリ16とメモリ17は、2つのテキスト文のそれぞれについて、木構造変換部4による変換結果を記憶する。メモリ18は、距離計算部8で求められたR〇木間或いはR木間の距離を記憶する。

なお、これらのメモリ10～18を1つにまとめた構成や、或いは、これらのメモリ10～18を設けないような構成が用いられてもよい。

## 【0032】

形態素解析部2は、メモリ10とメモリ11に格納されている2つのテキスト文の形態素及び属性を抽出し、それぞれのテキスト文の解析結果をそれぞれのメモリ12、メモリ13に記憶させる。

意味解析部3は、メモリ12とメモリ13に記録されている形態素解析の結果を入力し、テキスト文の意味を解析することを行い、それぞれのテキスト文の解析結果をそれぞれのメモリ14、メモリ15に格納する。

## 【0033】

木構造変換部4は、メモリ14とメモリ15に格納されている意味解析の結果をR〇木或いはR木に変換し、変換したR〇木或いはR木の頂点にテキスト文に出現した単語（単語の属性を含む）情報を記憶させるとともに辺にテキスト文に出現した関連する格情報を記憶させる。

また、木構造変換部4は、それぞれのテキスト文について変換した結果をそれ

ぞれのメモリ 1 6、メモリ 1 7に格納する。

【 0 0 3 4 】

単語写像重み計算部 5 は、R O 木間或いは R 木間の距離を求めるときに必要な単語置換重み、単語脱落重み、単語挿入重みを求めて、距離計算部 7 に提供する。

格写像重み計算部 6 は、R O 木間或いは R 木間の距離を求めるときに必要な格置換重み、格脱落重み、格挿入重みを求めて、距離計算部 7 に提供する。

【 0 0 3 5 】

距離計算部 7 は、メモリ 1 6 とメモリ 1 7 に格納されている 2 つの R O 木間或いは R 木間の距離を求め、その結果をメモリ 1 8 に記憶させる。

意味内容比較部 8 は、メモリ 1 8 に記憶されている R O 木間或いは R 木間の距離を利用して、テキスト文間の距離を求め、その結果を記憶部 9 に格納する。

【 0 0 3 6 】

次に、本発明に係るテキスト文間の意味内容を比較する距離を計算する装置及び方法の適用例として、情報端末装置に適用した場合の装置構成例を示す。

図 2 には、本発明に係るテキスト文間の意味内容を比較する距離を計算する方法を情報端末装置に適用した場合の装置構成例を示してある。

同図に示した情報端末装置 2 0 は、外部記憶装置 2 1 と、キーボード 2 2 と、ディスプレイ 2 3 と、プロセッサ部 2 4 から構成されている。プロセッサ部 2 4 には、テキスト文間の距離を求めるモジュール 2 5 が備えられている。

【 0 0 3 7 】

外部記憶装置 2 1 は、入力されたテキスト文のデータや、単語写像重みを求めるために用いられる単語特徴量辞書或いはシソーラス辞書や、格写像重みを求めるために用いられる重み辞書などや、求められたテキスト文間の距離の結果や、ソフトウェアなどを格納し、また、計算に用いられる記憶空間として機能する。ここで、単語特徴量辞書やシソーラス辞書や重み辞書などは、例えば、予め作成され或いは既存のものが用意される。また、具体的に、外部記憶装置 2 1 としては、例えばハードディスクなどで構成することができる。

【 0 0 3 8 】

キーボード 22 は、ユーザが操作を指示するための入力装置である。なお、他の入力装置が付加されていてもよい。

ディスプレイ 23 は、ユーザに対するメッセージやテキスト文のデータや、解析結果や、距離の計算結果などを表示するための出力装置である。なお、他の出力装置が付加されてもよい。

#### 【 0 0 3 9 】

プロセッサ部 24 は、外部記憶装置 21 に格納されているソフトウェアなどに従って、実際の処理を行う。具体的に、プロセッサ部 24 としては、例えば、マイクロプロセッサや、パーソナルコンピュータなどのコンピュータシステムで構成することができる。そして、上記図 1 に示した形態素解析部 2 や、意味解析部 3 や、木構造変換部 4 や、単語写像重み計算部 5 や、格写像重み計算部 6 や、距離計算部 7 や、意味内容比較部 8 は、このプロセッサ部 24 の上で動作するソフトウェアによって構成することができる。

#### 【 0 0 4 0 】

次に、本発明の一実施例に係るテキスト文間の意味内容の相違を比較する装置の動作を更に詳細に説明する。

外部記憶装置 1 には、テキスト文のデータを格納している。外部記憶装置 1 から 2 つのテキスト文のデータを読み出し、メモリ 10 とメモリ 11 にそれぞれ記憶させる。形態素解析部 2 は、メモリ 10 とメモリ 11 に記憶しているテキスト文の形態素を抽出し、その結果をそれぞれメモリ 12 とメモリ 13 に格納させる。

#### 【 0 0 4 1 】

ここで、形態素解析ツールとしては、公表された任意のものを利用することができ、例えば、奈良先端技術大学院大学の松本研究室により公表された“茶筌”形態素解析ツールを用いることができる。

また、図 3 には、テキスト文「先生は生徒に英語を教える」についての形態素解析の解析結果を示してある。

#### 【 0 0 4 2 】

意味解析部 3 は、メモリ 12 とメモリ 13 に記憶された形態素解析の結果を入

力し、テキスト文の構文や、係り受け関係（格）や、テキスト文の深層構造などを解析し、解析した結果をそれぞれメモリ14とメモリ15に格納する。

ここで、意味解析ツールとしては、公表された任意の意味解析ツールを利用することができ、例えば、非特許文献1に記載された方法を用いることができる（非特許文献1参照。）。

#### 【0043】

木構造変換部4は、メモリ14とメモリ15に記憶された解析結果を入力し、テキスト文を木構造へ変換して、変換した木構造をそれぞれメモリ16とメモリ17に格納する。

図4には、テキスト文「先生は生徒に英語を教える」についての意味解析の解析結果を木構造の形に書き換えたものを示してある。単語情報として、「先生」、「英語」、「生徒」、「に」、「教える」が各頂点に格納されており、格情報として、「先生」と「教える」の間の「SUBJ」、「英語」と「教える」の間の「OBJ」、「生徒」と「に」の間の「OBJ」、「に」と「教える」の間の「OBL」が各辺に格納されている。

#### 【0044】

上記図4において、格情報として、SUBJ（主格）、OBJ（目的格）、OBL（任意格）を示してある。また、格情報として、ADJUNCT（付加格）などを用いることもできる。

なお、本例では、OBLについては、格助詞と、言い換え可能な格助詞の数だけ変数を用意している。例えば、「彼は京都【に／へ】行った。」の場合、「に」と「へ」が言い換え可能なので、この変数名をOBL\_ni-heとする。

#### 【0045】

単語写像重み計算部5は、単語置換重み、単語脱落重み、単語挿入重みを求めて、距離計算部7に提供する。

単語置換重みとしては、定数と設定する態様や、単語間の距離を用いて設定する態様を用いることができる。前者の態様では、2つの単語が同じ単語である場合には単語置換重みをゼロと設定し、そうではない場合には単語置換重みを正の定数と設定する。後者の態様では、単語置換重み計算部5は、2つの単語間の距

離を求め、その距離の値を単語置換重みとして、距離計算部 7 に提供する。

#### 【0046】

ここで、単語間の距離を求める方法としては、公開された任意の方法を利用することができる。例えば、統計的な方法や、シソーラス辞書を用いた方法や、ニューラルネットを用いた方法がある。統計的な手法としては、例えば、非特許文献 2 に記載された TF/IDF 方法により求めることができる（非特許文献 2 参照。）。シソーラス辞書を用いた手法としては、例えば、2 つの単語が属している概念間の最短道の長さを単語間の距離とすることができる。ニューラルネットワークを用いた手法としては、例えば、非特許文献 3 に記載された方法を利用することができる（非特許文献 3 参照。）。また、他の公開された方法を利用することもできる。

#### 【0047】

単語脱落重みとしては、定数と設定する態様や、単語の品詞情報によって単語脱落重みを設定する態様を用いることができる。後者の態様では、単語の品詞に重みを付け、単語脱落重みを品詞重みと定数との積と設定する。品詞重みの設定としては、例えば、重要な役割を有する品詞に大きな重みを付与する仕方を用いることが好ましく、一例として、動詞の重みが一番重たく、形容動詞、名詞、副詞、形容詞などの順で品詞重みを軽くするように設定することができる。また、他の順番で品詞重みを設定することもできる。

#### 【0048】

単語挿入重みとしては、定数と設定する態様や、単語の品詞情報によって単語挿入重みを設定する態様を用いることができる。後者の態様では、単語の品詞に重みを付け、単語挿入重みを品詞重みと定数との積と設定する。品詞重みの設定としては、上記した単語脱落重みに関して述べた品詞重みの設定方法と同様な方法で設定することができ、また、異なる方法で設定することもできる。

#### 【0049】

格置換重み計算部 6 は、格置換重み、格脱落重み、格挿入重みを求めて、距離計算部 7 に提供する。

格置換重みとしては、定数と設定する態様や、格の間の距離を用いて設定する

態様を用いることができる。前者の態様では、2つの格が同じ格である場合には格置換重みをゼロと設定し、そうではない場合には各置換重みを正の定数と設定する。後者の態様では、格置換重み計算部6は、2つの格間の距離を求め、その距離の値を格置換重みとして、距離計算部7に提供する。

#### 【0050】

ここで、格間の距離を求める方法の一例を示す。

まず、全ての格をその内容によって幾つかのカテゴリに分類する。なお、カテゴリの要素数は1以上である。

また、図5に示されるような格カテゴリ間の距離の表を用意しておく。同図に示される表では、複数であるm個の格カテゴリの全ての組み合わせについて、格カテゴリ間の距離（距離11～距離mm）が設定されている。

次に、与えられた2つの格情報により特定される2つの格が属している格カテゴリをそれぞれ求め、上記図5に示される格カテゴリ間の距離表を用いて当該求められた2つの格カテゴリ間の距離値を求め、当該求められた距離値を2つの格間の距離とする。

なお、格間の距離を求める方法として、他の方法が用いられてもよい。

#### 【0051】

格脱落重みとしては、定数と設定する態様や、格の種類によって格脱落重みを設定する態様を用いることができる。後者の態様では、格に重みを付け、格脱落重みを格重みと定数との積と設定する。格重みの設定としては、例えば、SUBJの重みが一番重たく、OBJ、OBL、ADJUNCTなどの順で重みを軽くするように設定することができる。また、他の順番で格重みを設定することもできる。

#### 【0052】

格挿入重みとしては、定数と設定する態様や、格の種類によって格挿入重みを設定する態様を用いることができる。後者の態様では、格に重みを付け、格挿入重みを格重みと定数との積と設定する。格重みの設定としては、上記した格脱落重みに関して述べた格重みの設定方法と同様な方法で設定することができ、また、異なる方法で設定することもできる。

## 【0053】

距離計算部7は、メモリ16とメモリ17に記憶されたR〇木間或いはR木間の距離を求め、その結果をメモリ18に格納する。

ここで、頂点と辺の対応関係に基づくR〇木間の距離は、例えば、特願2002-071273号に記載された方法で求めることができる。また、頂点と辺の対応関係に基づくR木間の距離は、例えば、特願2002-071274号に記載された方法で求めることができる。

## 【0054】

次に、上記した特願2002-071273号に記載されたR〇木間の距離を求める方法や、上記した特願2002-071274号に記載されたR木間の距離を求める方法を示す。

まず、木の間の距離を記述するために、関連する記号を定義する。

R〇木或いはR木 $T_a$ の頂点 $x$ を根とする部分木を $T_a(x)$ で表す。

部分木 $T_a(x)$ の頂点の集合を $V_a(x)$ で表す。

部分木 $T_a(x)$ の辺の集合を $E_a(x)$ で表す。

頂点 $x$ の子供を $x_1, x_2, \dots, x_m$ とし、頂点 $x$ の子供の集合を $Ch(x)$ で表す。

また、本明細書では、次のように定義する。

## 【0055】

## 【数1】

部分木 $T_a(x)$ と辺 $\tilde{x}$ からなる部分を $\tilde{T}_a(x)$ で表し、 $\tilde{T}_a(x)$ も部分木と呼ぶ。

部分木 $\tilde{T}_a(x)$ の頂点の集合を $\tilde{V}_a(x)$ で表し、部分木 $\tilde{T}_a(x)$ の辺の集合を $\tilde{E}_a(x)$ で表す。この場合、 $V_a(x) = \tilde{V}_a(x)$ となり、 $E_a(x) \cup \{\tilde{x}\} = \tilde{E}_a(x)$ となる。

部分木 $\tilde{T}_a(x_1), \tilde{T}_a(x_2), \dots, \tilde{T}_a(x_m)$ からなる部分を森と呼び、森を $\tilde{F}_a(x)$ で表す。

## 【0056】

また、上述したテキスト文の木構造表現法と変換方法から分かるように、頂点  $x$ 、 $y$  はテキスト文に出現している単語（単語の属性を含む）を表している。また、関数  $\delta(x, y)$  は頂点の置換重みを表すとし、単語の置換重みで求めることができる。また、 $q(x)$  は頂点  $x$  の挿入重みを表すとし、単語の挿入重みで求めることができる。また、 $r(x)$  は頂点  $x$  の脱落重みを表すとし、単語の脱落重みで求めることができる。

また、辺に関して、次のように定義する。

【0057】

【数2】

上述したテキスト文の木構造表現法と変換方法から分かるように、辺  $\tilde{x}$ 、 $\tilde{y}$  はテキスト文の単語間の係り受け関係（格）の情報を表している。また、関数  $\delta(\tilde{x}, \tilde{y})$  は辺の置換重みを表すとし、格の置換重みで求めることができる。また、 $q(\tilde{x})$  は辺  $\tilde{x}$  の挿入重みを表すとし、格の挿入重みで求めることができる。また、 $r(\tilde{x})$  は辺  $\tilde{x}$  の脱落重みを表すとし、格の脱落重みで求めることができる。

【0058】

初めに、上記した特願 2002-071273 号に記載された頂点と辺の対応関係に基づく RO 木間の距離を求める方法を示す。

まず、RO 木の根から深さ優先順で頂点と辺に番号を付ける。大きい番号を根とする RO 木の部分から小さい番号を根とする RO 木の部分の順で部分木間の距離を求めて、最後に全体的な RO 木間の距離を求める。

【0059】

図 6 (a) ~ 図 6 (d) には、例えば RO 木である 2 つの部分木を 4 つの態様について示してある。

RO 木である部分木間の距離及び順序がある森間の距離に関して、次のように定義する。

【0060】



## 【数 3】

図 6 (a) に示される R O 木である 2 つの部分木  $T_a(x)$ 、 $T_b(y)$  の間の距離を  $D(T_a(x), T_b(y))$  で表す。

図 6 (b) に示される R O 木である 2 つの部分木  $\tilde{T}_a(x)$ 、 $\tilde{T}_b(y)$  の間の距離を  $D(\tilde{T}_a(x), \tilde{T}_b(y))$  で表す。

図 6 (c) に示される R O 木である 2 つの部分木  $\tilde{T}_a(x)$ 、 $T_b(y)$  の間の距離を  $D(\tilde{T}_a(x), T_b(y))$  で表す。

図 6 (d) に示される R O 木である 2 つの部分木  $T_a(x)$ 、 $\tilde{T}_b(y)$  の間の距離を  $D(T_a(x), \tilde{T}_b(y))$  で表す。

また、順序がある森間の距離  $D(\tilde{F}_a(x), \tilde{F}_b(y))$ 、全ての部分木間の距離  $D(T_a(x_i), T_b(y_j))$ 、 $D(\tilde{T}_a(x_i), \tilde{T}_b(y_j))$ 、 $D(\tilde{T}_a(x_i), T_b(y_j))$ 、 $D(T_a(x_i), \tilde{T}_b(y_j))$ 、 $D(T_a(x), T_b(y_j))$ 、 $D(\tilde{T}_a(x), \tilde{T}_b(y_j))$ 、 $D(\tilde{T}_a(x), T_b(y_j))$ 、 $D(T_a(x), \tilde{T}_b(y_j))$  は既に求められたものとする。

## 【0061】

式 1 ～ 式 4 を用いて、同図 (a) ～ 同図 (d) のそれぞれに示される 2 つの R O 木の間の距離を求めることができる。ここで、式 1 ～ 式 4 中で記号 “A - B” は集合 A から集合 B の全ての要素を取り除く関数を表す。

## 【0062】

## 【数 4】

$$D(T_a(x), T_b(y)) = \min \left\{ \begin{array}{l} \delta(x, y) + D(\tilde{F}_a(x), \tilde{F}_b(y)), \\ \min_{x_i \in \mathcal{H}(x)} \left\{ D(\tilde{T}_a(x_i), T_b(y)) + \sum r(k) | k \in (V_a(x) - V_a(x_i)) + \sum r(\tilde{k}) | \tilde{k} \in (E(x) - \tilde{E}(x_i)) \right\}, \\ \min_{y_j \in \mathcal{H}(y)} \left\{ D(T_a(x), \tilde{T}_b(y_j)) + \sum q(k) | k \in (V_b(y) - V_b(y_j)) + \sum q(\tilde{k}) | \tilde{k} \in (E(y) - \tilde{E}(y_j)) \right\} \end{array} \right\}$$

.. (式 1)

## 【0063】

【数 5】

$$D(\tilde{T}_a(x), \tilde{T}_b(y)) = \min \left\{ \begin{array}{l} \delta(x, y) + \delta(\tilde{x}, \tilde{y}) + D(\tilde{F}_a(x), \tilde{F}_b(y)), \\ \min_{x \in CH(x)} \left\{ \min \{ D(T_a(x), T_b(y)) + \delta(\tilde{x}, \tilde{y}), D(\tilde{T}_a(x), \tilde{T}_b(y)) \} \right. \\ \quad \left. + \sum r(k) | k \in (\tilde{V}_a(x) - \tilde{V}_a(x)) + \sum r(\tilde{k}) | \tilde{k} \in (\tilde{E}_a(x) - \tilde{E}_a(x)) \right\}, \\ \min_{y \in CH(y)} \left\{ \min \{ D(T_a(x), T_b(y)) + \delta(\tilde{x}, \tilde{y}), D(\tilde{T}_a(x), \tilde{T}_b(y)) \} \right. \\ \quad \left. + \sum q(k) | k \in (\tilde{V}_b(y) - \tilde{V}_b(y)) + \sum q(\tilde{k}) | \tilde{k} \in (\tilde{E}_b(y) - \tilde{E}_b(y)) \right\} \end{array} \right\}$$

.. (式 2)

【0064】

【数 6】

$$D(\tilde{T}_a(x), \tilde{T}_b(y)) = \min \left\{ \begin{array}{l} \delta(x, y) + \tilde{r} + D(\tilde{F}_a(x), \tilde{F}_b(y)), \\ \min_{x \in CH(x)} \left\{ D(\tilde{T}_a(x), \tilde{T}_b(y)) + \sum r(k) | k \in (V_a(x) - V_a(x)) + \sum r(\tilde{k}) | \tilde{k} \in (\tilde{E}_a(x) - \tilde{E}_a(x)) \right\}, \\ \min_{y \in CH(y)} \left\{ D(\tilde{T}_a(x), \tilde{T}_b(y)) + \sum q(k) | k \in (V_b(y) - V_b(y)) + \sum q(\tilde{k}) | \tilde{k} \in (\tilde{E}_b(y) - \tilde{E}_b(y)) \right\} \end{array} \right\}$$

.. (式 3)

【0065】

【数 7】

$$D(\tilde{T}_a(x), \tilde{T}_b(y)) = \min \left\{ \begin{array}{l} \delta(x, y) + \tilde{q} + D(\tilde{F}_a(x), \tilde{F}_b(y)), \\ \min_{x \in CH(x)} \left\{ D(\tilde{T}_a(x), \tilde{T}_b(y)) + \sum r(k) | k \in (V_a(x) - V_a(x)) + \sum r(\tilde{k}) | \tilde{k} \in (\tilde{E}_a(x) - \tilde{E}_a(x)) \right\}, \\ \min_{y \in CH(y)} \left\{ D(\tilde{T}_a(x), \tilde{T}_b(y)) + \sum q(k) | k \in (V_b(y) - V_b(y)) + \sum q(\tilde{k}) | \tilde{k} \in (\tilde{E}_b(y) - \tilde{E}_b(y)) \right\} \end{array} \right\}$$

.. (式 4)

【0066】

図 7 には、例えば順序がある 2 つの森を示してある。式 5 を用いて、これら 2 つの森の間の距離を求めることができる。ここで、記号  $|A|$  は集合  $A$  の要素数

を表す。

【0067】

【数8】

同図に示される2つの順序がある森  $\tilde{F}_a(x)$ 、 $\tilde{F}_b(x)$  の間の距離  $D(\tilde{F}_a(x), \tilde{F}_b(y))$  が次のように求められる。

(5-1) 境界条件 ( $1 \leq i \leq |Ch(x)|$ ,  $1 \leq j \leq |Ch(y)|$ )

$$d1(0,0)=0;$$

$$d1(i,0)=d1(i-1,0)+\sum r(k)|k \in A(x)+\sum r(\tilde{k})|\tilde{k} \in \tilde{E}(x);$$

$$d1(0,j)=d1(0,j-1)+\sum q(k)|k \in A(y)+\sum q(\tilde{k})|\tilde{k} \in \tilde{E}(y);$$

(5-2)  $d1(i,j)$  の計算 ( $1 \leq i \leq |Ch(x)|$ ,  $1 \leq j \leq |Ch(y)|$ )

$$d1(i,j)=\min \left\{ \begin{array}{l} d1(i-1,j-1)+D(\tilde{T}_a(x), \tilde{T}_b(y)), \\ d1(i,j-1)+\sum q(k)|k \in A(y)+\sum q(\tilde{k})|\tilde{k} \in \tilde{E}(y), \\ d1(i-1,j)+\sum r(k)|k \in A(x)+\sum r(\tilde{k})|\tilde{k} \in \tilde{E}(x) \end{array} \right\};$$

$$(5-3) \quad D(\tilde{F}_a(x), \tilde{F}_b(y)) = d1(|Ch(x)|, |Ch(y)|).$$

.. (式5)

【0068】

なお、上記式1について、頂点  $x$  が葉 ( $Ch(x) = \text{NULL}$ : 空集合) である場合には、明らかに、上記式1の右側の第2項を計算する必要がないため、式6を用いて距離  $D(T_a(x), T_b(y))$  を求めることができる。

また、上記式1について、頂点  $y$  が葉 ( $Ch(y) = \text{NULL}$ : 空集合) である場合には、明らかに、上記式1の右側の第3項を計算する必要がないため、式7を用いて距離  $D(T_a(x), T_b(y))$  を求めることができる。

【0069】

【数 9】

$$D(T(x), T(y)) = \min \left\{ \begin{aligned} &\delta(x, y) + D(\tilde{F}(x), \tilde{F}(y)), \\ &\min_{y \in \mathcal{H}(y)} \left\{ D(T(x), \tilde{T}(y)) + \sum q(k) | k \in V(y) - V(y) \right\} + \sum q(\tilde{k}) | \tilde{k} \in (E(y) - \tilde{E}(y)) \right\} \end{aligned} \right\}$$

.. (式 6)

【0 0 7 0】

【数 1 0】

$$D(T(x), T(y)) = \min \left\{ \begin{aligned} &\delta(x, y) + D(\tilde{F}(x), \tilde{F}(y)), \\ &\min_{x \in \mathcal{H}(x)} \left\{ D(\tilde{T}(x), T(y)) + \sum r(k) | k \in V(x) - V(x) \right\} + \sum r(\tilde{k}) | \tilde{k} \in (E(x) - \tilde{E}(x)) \right\} \end{aligned} \right\}$$

.. (式 7)

【0 0 7 1】

同様に、上記式 2 について、頂点  $x$  が葉 ( $\text{Ch}(x) = \text{NULL}$  : 空集合) である場合には、明らかに、上記式 2 の右側の第 2 項を計算する必要がないため、式 8 を用いて距離を求めることができる。

また、上記式 2 について、頂点  $y$  が葉 ( $\text{Ch}(y) = \text{NULL}$  : 空集合) である場合には、明らかに、上記式 2 の右側の第 3 項を計算する必要がないため、式 9 を用いて距離を求めることができる。

【0 0 7 2】

【数 1 1】

$$D(\tilde{T}(x), \tilde{T}(y)) = \min \left\{ \begin{aligned} &\delta(x, y) + \delta(\tilde{x}, \tilde{y}) + D(\tilde{F}(x), \tilde{F}(y)), \\ &\min_{y \in \mathcal{H}(y)} \left\{ \min \left\{ D(T(x), T(y)) + \delta(\tilde{x}, \tilde{y}), D(\tilde{T}(x), \tilde{T}(y)) \right\} \right. \\ &\quad \left. + \sum q(k) | k \in \tilde{V}(y) - \tilde{V}(y) \right\} + \sum q(\tilde{k}) | \tilde{k} \in (\tilde{E}(y) - \tilde{E}(y)) \right\} \end{aligned} \right\}$$

.. (式 8)

【0 0 7 3】

【数 1 2】

$$D(\tilde{T}_a(x), \tilde{T}_b(y)) = \min \left\{ \begin{array}{l} \delta(x, y) + \delta(\tilde{x}, \tilde{y}) + D(\tilde{F}_a(x), \tilde{F}_b(y)), \\ \min_{x \in Ch(x)} \left\{ \min \{ D(T_a(x), T_b(y)) + \delta(\tilde{x}, \tilde{y}), D(\tilde{T}_a(x), \tilde{T}_b(y)) \} \right. \right. \\ \left. \left. + \sum r(k) | k \in (\tilde{V}_a(x) - \tilde{V}_b(x)) + \sum r(\tilde{k}) | \tilde{k} \in (\tilde{E}_a(x) - \tilde{E}_b(x)) \right\} \right\} \end{array} \right.$$

.. (式 9)

【0074】

同様に、上記式 3 について、頂点  $x$  が葉 ( $Ch(x) = \text{NULL}$  : 空集合) である場合には、明らかに、上記式 3 の右側の第 2 項を計算する必要がないため、式 10 を用いて距離を求めることができる。

また、上記式 3 について、頂点  $y$  が葉 ( $Ch(y) = \text{NULL}$  : 空集合) である場合には、明らかに、上記式 3 の右側の第 3 項を計算する必要がないため、式 11 を用いて距離を求めることができる。

【0075】

【数 1 3】

$$D(\tilde{T}_a(x), T_b(y)) = \min \left\{ \begin{array}{l} \delta(x, y) + \tilde{r} + D(\tilde{F}_a(x), \tilde{F}_b(y)), \\ \min_{y \in Ch(y)} \left\{ D(\tilde{T}_a(x), \tilde{T}_b(y)) + \sum q(k) | k \in (V_a(y) - V_b(y)) + \sum q(\tilde{k}) | \tilde{k} \in (E_a(y) - E_b(y)) \right\} \right\} \end{array} \right.$$

.. (式 10)

【0076】

【数 1 4】

$$D(\tilde{T}_a(x), T_b(y)) = \min \left\{ \begin{array}{l} \delta(x, y) + \tilde{r} + D(\tilde{F}_a(x), \tilde{F}_b(y)), \\ \min_{x \in Ch(x)} \left\{ D(\tilde{T}_a(x), T_b(y)) + \sum r(k) | k \in (V_a(x) - V_b(x)) + \sum r(\tilde{k}) | \tilde{k} \in (\tilde{E}_a(x) - \tilde{E}_b(x)) \right\} \right\} \end{array} \right.$$

.. (式 11)

【0077】

同様に、上記式 4 について、頂点  $x$  が葉 ( $Ch(x) = NUL L$  : 空集合) である場合には、明らかに、上記式 4 の右側の第 2 項を計算する必要がないため、式 12 を用いて距離を求めることができる。

また、上記式 4 について、頂点  $y$  が葉 ( $Ch(y) = NUL L$  : 空集合) である場合には、明らかに、上記式 4 の右側の第 3 項を計算する必要がないため、式 13 を用いて距離を求めることができる。

【0078】

【数 15】

$$D(T(x), \tilde{T}(y)) = \min \left\{ \begin{array}{l} \delta(x, y) + \tilde{q} + D(\tilde{T}(x), \tilde{T}(y)), \\ \min_{y \in \mathcal{H}(y)} \left\{ D(T(x), \tilde{T}(y)) + \sum_{k \in V(y) - V(y)} q(k) + \sum_{\tilde{k} \in E(y) - \tilde{E}(y)} q(\tilde{k}) \right\} \end{array} \right.$$

.. (式 12)

【0079】

【数 16】

$$D(T(x), \tilde{T}(y)) = \min \left\{ \begin{array}{l} \delta(x, y) + \tilde{q} + D(\tilde{T}(x), \tilde{T}(y)), \\ \min_{x \in \mathcal{H}(x)} \left\{ D(\tilde{T}(x), \tilde{T}(y)) + \sum_{k \in V(x) - V(x)} r(k) + \sum_{\tilde{k} \in E(x) - \tilde{E}(x)} r(\tilde{k}) \right\} \end{array} \right.$$

.. (式 13)

【0080】

次に、上記した特願 2002-071274 号に記載された頂点と辺の対応関係に基づく R 木間の距離を求める方法を示す。

まず、R 木の根から深さ優先順で頂点と辺に番号を付ける。大きい番号を根とする R 木の部分から小さい番号を根とする R 木の部分の順で部分木間の距離を求めて、最後に全体的な R 木間の距離を求める。

【0081】

## 【数 1 7】

本方法では、上記図 6 (a) に示されるような例えば R 木である 2 つの部分木  $T_a(x)$ 、 $T_b(y)$  の間の距離  $D(T_a(x), T_b(y))$  を上記式 1 で求めることができ、上記図 6 (b) に示されるような例えば R 木である 2 つの部分木  $\tilde{T}_a(x)$ 、 $\tilde{T}_b(y)$  の間の距離  $D(\tilde{T}_a(x), \tilde{T}_b(y))$  を上記式 2 で求めることができ、上記図 6 (c) に示されるような例えば R 木である 2 つの部分木  $\tilde{T}_a(x)$ 、 $T_b(y)$  の間の距離  $D(\tilde{T}_a(x), T_b(y))$  を上記式 3 で求めることができ、上記図 6 (d) に示されるような例えば R 木である 2 つの部分木  $T_a(x)$ 、 $\tilde{T}_b(y)$  の間の距離  $D(T_a(x), \tilde{T}_b(y))$  を上記式 4 で求めることができる。

ここで、順序がない森間の距離  $D(\tilde{F}_a(x), \tilde{F}_b(y))$ 、全ての部分木間の距離  $D(T_a(x_i), T_b(y))$ 、 $D(\tilde{T}_a(x_i), \tilde{T}_b(y))$ 、 $D(\tilde{T}_a(x_i), T_b(y))$ 、 $D(T_a(x_i), \tilde{T}_b(y))$ 、 $D(T_a(x), T_b(y_j))$ 、 $D(\tilde{T}_a(x), \tilde{T}_b(y_j))$ 、 $D(\tilde{T}_a(x), T_b(y_j))$ 、 $D(T_a(x), \tilde{T}_b(y_j))$  は既に求められたものとする。

## 【0 0 8 2】

式 1 4 を用いて、上記図 7 に示されるような例えば順序がない 2 つの森の間の距離を求めることができる。

## 【0 0 8 3】

【数 18】

同図に示されるような2つの順序がない森  $\tilde{F}_a(x)$ 、 $\tilde{F}_b(x)$  の間の距離  $D(\tilde{F}_a(x), \tilde{F}_b(y))$  が次のように求められる。

$$D(\tilde{F}_a(x), \tilde{F}_b(y)) = \sum_{x \in Ch(x)} \left( \sum r(k) | k \in \tilde{F}_a(x) + \sum r(\tilde{k}) | \tilde{k} \in \tilde{E}(x) \right) \\ + \sum_{y \in Ch(y)} \left( \sum q(k) | k \in \tilde{F}_b(y) + \sum q(\tilde{k}) | \tilde{k} \in \tilde{E}(y) \right) - W(M_{max})$$

.. (式14)

【0084】

ここで、上記式14中の  $W(M_{max})$  は、図8に示すような2部グラフ  $G(X, Y, E)$  の最大マッチングの重みである。

また、2部グラフ  $G(X, Y, E)$  の頂点  $x_i (\in X)$  と頂点  $y_j (\in Y)$  の間の辺  $e(x_i, y_j)$  の重み  $w(e(x_i, y_j))$  を式15のように設定する。2部グラフ  $G(X, Y, E)$  の最大マッチングの重みは、辺  $e(x_i, y_j)$  の重み  $w(e(x_i, y_j))$  の和の最大値に相当する。

【0085】

【数19】

$$w(e(x, y)) = \sum r(k) | k \in \tilde{F}_a(x) + \sum r(\tilde{k}) | \tilde{k} \in \tilde{E}(x) + \sum q(k) | k \in \tilde{F}_b(y) + \sum q(\tilde{k}) | \tilde{k} \in \tilde{E}(y) - D(\tilde{T}_a(x), \tilde{T}_b(y))$$

.. (式15)

なお、2部グラフ  $G(X, Y, E)$  の頂点  $x_i (\in X)$  は順序がない森  $\tilde{F}_a(x)$  を構成する部分木  $\tilde{T}_a(x_i)$  ( $x_i \in Ch(x)$ ) を表し、2部グラフ  $G(X, Y, E)$  の頂点  $y_j (\in Y)$  は順序がない森  $\tilde{F}_b(y)$  を構成する部分木  $\tilde{T}_b(y_j)$  ( $y_j \in Ch(y)$ ) を表す。

【0086】

以上のような方法により、RO木間或いはR木間の距離  $D(T_a, T_b) = D$



$(T_a (x=1), T_b (y=1))$  を求めることができる。

次に、意味内容比較部 8 は、式 1 6 或いは式 1 7 を用いて、テキスト文間の距離を求める。

ここで、 $D(S_1, S_2)$  は文  $S_1$  と文  $S_2$  との間の距離を表し、木  $T_1$  は文  $S_1$  の木構造 (RO 木或いは R 木) を表し、木  $T_2$  は文  $S_2$  の木構造 (RO 木或いは R 木) を表し、 $D(T_1, T_2)$  は木  $T_1$  と木  $T_2$  との間の距離を表す。

【0087】

【数 2 0】

$$D(S_1, S_2) = D(T_1, T_2) \quad \dots (式 1 6)$$

【0088】

【数 2 1】

$$D(S_1, S_2) = \frac{D(T_1, T_2)}{|T_1| + |T_2|} \quad \dots (式 1 7)$$

【0089】

次に、具体的な例を用いて、本発明の一実施例に係るテキスト文の意味内容の比較装置及び比較方法の動作を説明する。

本発明の一実施例に係るテキスト文の意味内容の比較装置を用いて、文 A 「妻の花子は風邪を引きました」と文 B 「妻は風邪を引きました」の間の類似度 (或いは、相違度) を求める過程と結果を示す。本例では、単語と格の脱落重み及び挿入重みを 7 0 と設定し、単語間の置換重みを 1 0 0 と設定し、格間の置換重みを 1 0 0 と設定する。

【0090】

まず、文 A と文 B を形態素解析した後に、構文意味解析を行い、これにより、これら 2 つの文 A、B がそれぞれ図 9 (a)、図 9 (b) に示すような例えば根があり順序がある木  $T_A$ 、 $T_B$  に変換される。

次に、上記式 1 を用いて変換された 2 つの RO 木間の距離を計算し、最後に、

上記式 1 6 或いは上記式 1 7 を用いて 2 つのテキスト文 A、テキスト文 B の間の距離を求める。

#### 【 0 0 9 1 】

上記式 1 6 を用いた場合には、テキスト文 A、B 間の距離は  $D(A, B) = 140$  となり、上記式 1 7 を用いた場合には、テキスト文 A、B 間の距離は  $D(A, B) = 20 (= 140 / 7)$  となる。ここで、2 つの R O 木  $T_A$ 、 $T_B$  間の距離は  $D(T_A, T_B) = 140$  であり、2 つの R O 木  $T_A$ 、 $T_B$  の頂点の総数は 7 である。

#### 【 0 0 9 2 】

図 1 0 には、距離  $D(T_A, T_B)$  を与える R O 木間の写像の一つを示している。同図に示されるように、2 つの R O 木  $T_A$ 、 $T_B$  間の距離は、単語「花子」を脱落するのに必要な脱落重みである 7 0 と、格「ADJUNCT」を脱落するのに必要な脱落重みである 7 0 との和となっている。

#### 【 0 0 9 3 】

以上のように、本発明に係るテキスト文比較装置及び比較方法では、テキスト文を形態素解析し、意味解析を行い、解析されたテキスト文の全体の構文構造と意味をグラフ理論上の R O 木或いは R 木で表現し、つまり、テキスト文の全体の構文構造と意味を R O 木或いは R 木に変換し、テキスト文に出現した単語情報（単語の属性を含む）と単語間の係り受け関係（格）情報をそれぞれ R O 木或いは R 木の頂点と辺に格納し、頂点と辺の対応関係に基づく R O 木間或いは R 木間の距離をテキスト文間の意味内容の相違を測る距離に適用して、R O 木間或いは R 木間の距離を用いてテキスト文間の意味内容の相違を比較することにより、入力された 2 つのテキスト文間の意味内容を高精度に且つ実時間で求めることができる。

#### 【 0 0 9 4 】

具体的には、本発明では、テキスト文間の距離を、テキスト文間の単語情報の違い、格情報の違い、及び文全体の構造上の違いによって定義したため、本発明に係る距離関数は次の 3 つの良い性質を有している。（1）意味が似ている且つ構造が似ている 2 つの文間の距離が非常に小さく評価される。（2）意味が異な

る且つ構造が似ていない2つの文間の距離が非常に大きく評価される。(3)意味が異なるが、構造が似ている2つの文間の距離が単語情報の違いと格情報の違いによって評価される。これにより、2つのテキスト文間の距離を高精度に求めることができる。

## 【0095】

また、本例では、R O 木については、木の頂点の数  $n$  の2乗のオーダー ( $O(n^2)$ ) で計算することが可能であり、また、R 木については、木の頂点の数  $n$  の2乗と木の最大の子供の数  $m$  のオーダー ( $O(mn^2)$ ) で計算することが可能である。このように実時間での計算が可能である。

## 【0096】

ここで、本発明の構成としては、必ずしも以上に示したものに限られず、種々な構成が用いられてもよい。なお、本発明は、例えば本発明に係る方法を実現するためのプログラムなどとして提供することも可能である。

また、本発明の適用分野としては、必ずしも以上に示したものに限られず、本発明は、種々な分野に適用することが可能なものである。

## 【0097】

また、本発明において行われる各種の処理としては、例えばプロセッサやメモリ等を備えたハードウェア資源においてプロセッサがROM (Read Only Memory) に格納された制御プログラムを実行することにより制御される構成が用いられてもよく、また、例えば当該処理を実行するための各機能手段が独立したハードウェア回路として構成されてもよい。

また、本発明は上記の制御プログラムを格納したフロッピー (登録商標) ディスクやCD (Compact Disc) - ROM等のコンピュータにより読み取り可能な記録媒体や当該プログラム (自体) として把握することもでき、当該制御プログラムを記録媒体からコンピュータに入力してプロセッサに実行させることにより、本発明に係る処理を遂行させることができる。

## 【0098】

## 【発明の効果】

以上説明したように、本発明に係るテキスト文比較装置やテキスト文比較方法

によると、テキスト文の全体の構文と意味をグラフ理論上のR O木或いはR木で表現し、頂点と辺の対応関係に基づくR O木間の距離或いは頂点と辺の対応関係に基づくR木間の距離を用いてテキスト文間の意味内容の相違を比較することにより、入力された2つのテキスト文間の意味内容を高精度に且つ実時間で求めることができる。本発明により、例えば、ドキュメントの意味内容の比較や、意味内容によるドキュメントの分類だけではなく、ユーザの情報探索意図の理解なども可能になる。つまり、自然言語で表現したユーザの要求を、事前に学習して構築されたデータベースの中の内容と比較して、ユーザの情報探索意図を推定することができる。

#### 【図面の簡単な説明】

【図1】 本発明の一実施例に係るテキスト文間の意味内容の比較装置の構成例を示す図である。

【図2】 本発明に係るテキスト文間の意味内容の比較装置及び比較方法を情報端末装置に適用した場合の構成例を示す図である。

【図3】 形態素解析部による解析結果の一例を示す図である。

【図4】 木構造の表現の一例を示す図である。

【図5】 格カテゴリ間の距離の表（リスト）のデータ構造の一例を示す図である。

【図6】 R O木或いはR木からなる2つの部分木の一例を示す図である。

【図7】 R O木或いはR木からなる2つの森の一例を示す図である。

【図8】 2部グラフの一例を示す図である。

【図9】 文Aと文Bの木構造を示す図である。

【図10】 文Aと文BのR O木間の距離を与える写像の一例を示す図である。

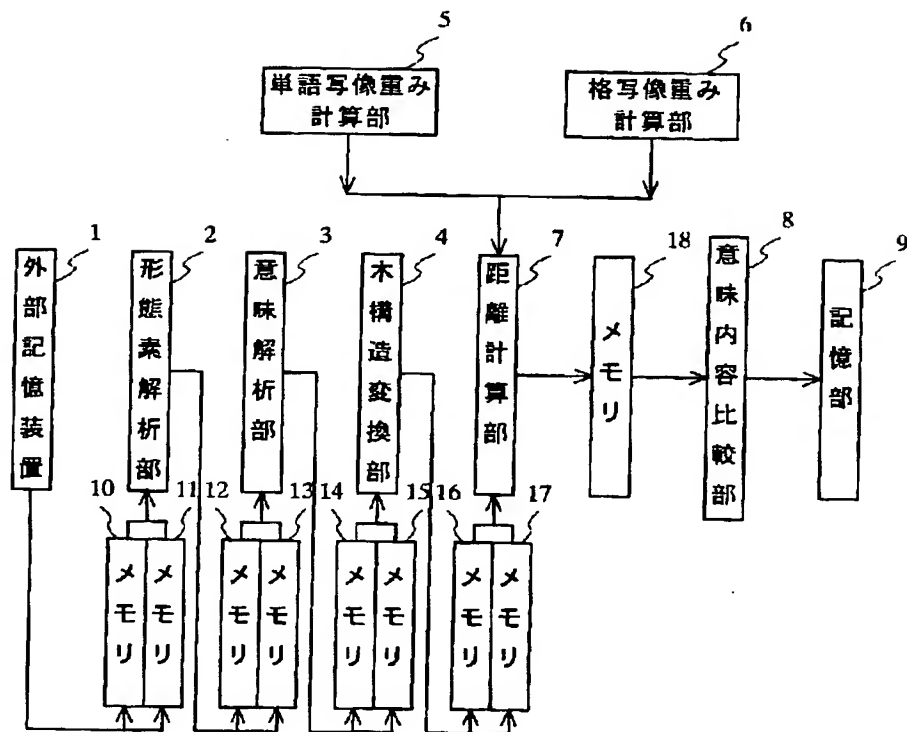
#### 【符号の説明】

- 1、21・・・外部記憶装置、 2・・・形態素解析部、 3・・・意味解析部、
- 4・・・木構造変換部、 5・・・単語写像重み計算部、
- 6・・・格写像重み計算部、 7・・・距離計算部、 8・・・意味内容比較部、
- 9・・・記憶部、 10～18・・・メモリ、 20・・・情報端末装置、

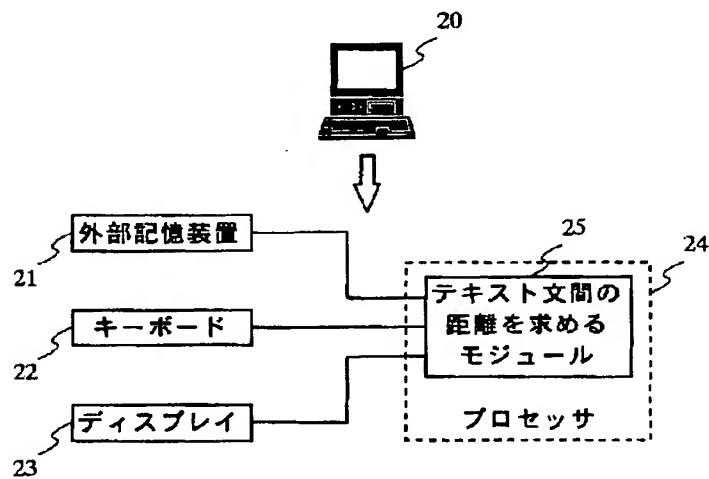
2 2 . . キーボード、 2 3 . . ディスプレイ、 2 4 . . プロセッサ部、  
2 5 . . モジュール、

【書類名】 図面

【図 1】



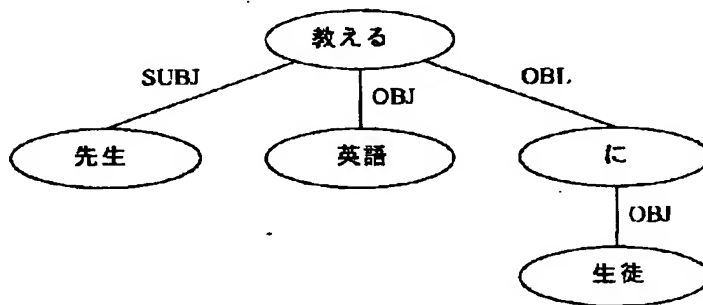
【図 2】



【図3】

先生	先生	センセイ	センセイ	名詞—一般
は	は	ハ	ワ	助詞—係助詞
生徒	生徒	セイト	セイト	名詞—一般
に	に	ニ	ニ	助詞—格助詞—一般
英語	英語	エイゴ	エイゴ	名詞—一般
を	を	ヲ	ヲ	助詞—格助詞—一般
教える	教える	オシエル	オシエル	動詞—自立 一段 基本形
EOS				

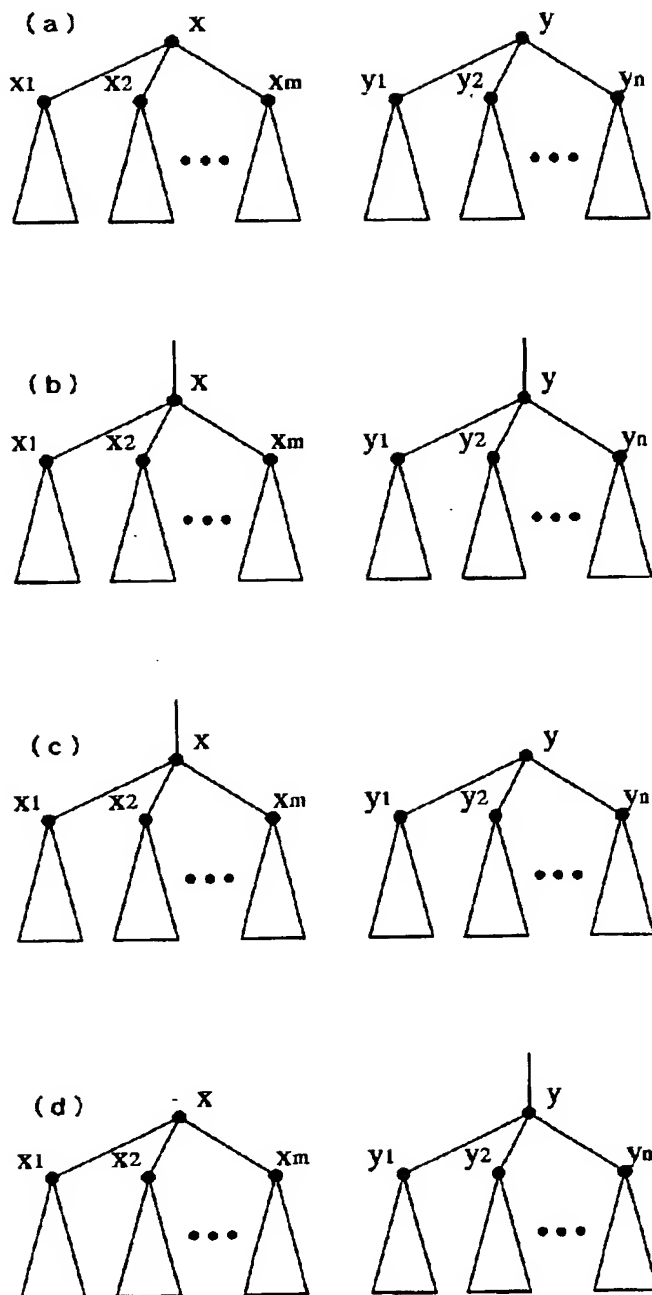
【図4】



【図5】

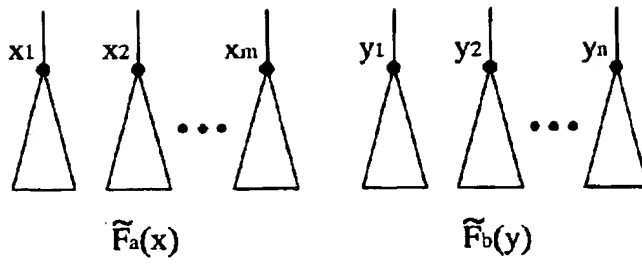
格カテゴリ1	格カテゴリ1	距離値11
格カテゴリ1	格カテゴリ2	距離値12
...	...	
格カテゴリ1	格カテゴリm	距離値1m
...	...	
格カテゴリm	格カテゴリ1	距離値m1
格カテゴリm	格カテゴリ2	距離値m2
...	...	
格カテゴリm	格カテゴリm	距離値mm

【図 6】

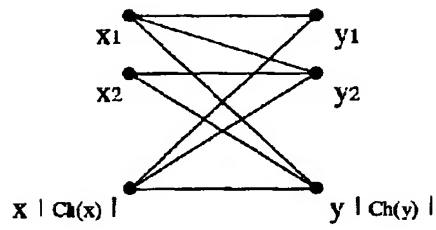




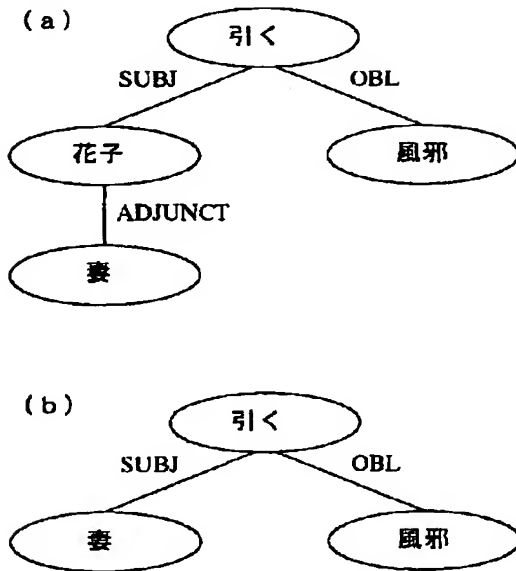
【図 7】



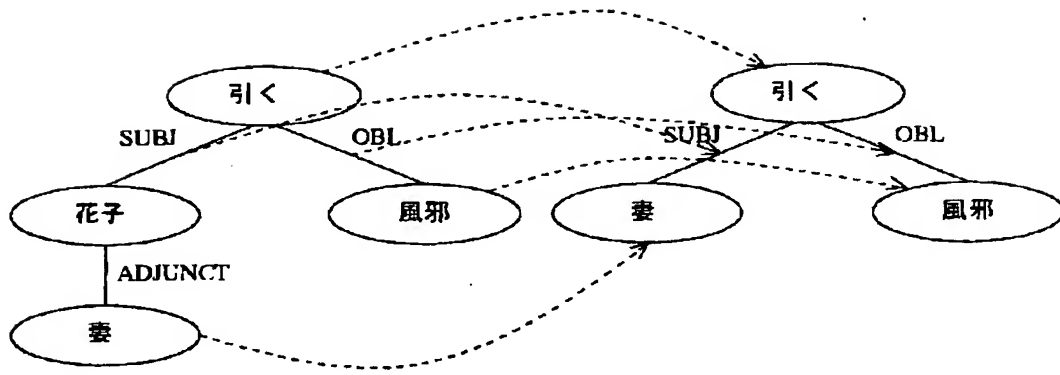
【図 8】



【図 9】



【図10】



【書類名】 要約書

【要約】

【課題】 高精度に実時間でテキスト文間の比較を行うテキスト文比較装置を提供する。

【解決手段】 木表現手段4が比較対象となるテキスト文をグラフ理論上の根がある木で表現し、情報付与手段4が木表現手段4により表現される木の各頂点に単語情報を付与して各辺に単語間の係り受け関係情報である格情報を付与し、木間距離定義手段7が頂点間の対応関係及び辺間の対応関係に基づく木の間の距離を定義し、木間距離取得手段7が比較対象となるテキスト文の木間について木間距離定義手段7により定義される木間の距離を求め、木間距離適用手段8が木間の距離をテキスト文間の相違を表す距離に適用し、木間距離取得手段8が木間距離適用手段8による適用に基づいて比較対象となるテキスト文間の距離を求める。

【選択図】 図1

出 願 人 履 歴 情 報

識別番号 [000005496]

1. 変更年月日	1996年 5月29日
[変更理由]	住所変更
住 所	東京都港区赤坂二丁目17番22号
氏 名	富士ゼロックス株式会社